includes two steps where the first step is designed to remove the hard mask layer **224** and the second step is designed to remove the silicon in the gate stacks in the device region **214** and the overlay region **212**. After the silicon in the gate stacks is removed, trenches result in the ILD layer **242** and are referred to as gate trenches.

In one embodiment, the first etch step to remove the hard mask layer **242** may include phosphoric acid (H_3PO_4) solution, hydrofluoric acid (HF), or buffered HF if the hard mask layer **224** include silicon nitride. In another embodiment, the etching process used to remove the polysilcion or amorphous silicon of the gate stacks may be dry etching, wet etching or combinations thereof. In one example, an etching solution including HNO_3, H_2O and HF, or NH_4OH solution, may be used to remove polysilicon (or amorphous silicon). In another example, chlorine (Cl)-based plasma may be used to selectively remove the polysilicon.

After the formation of the gate trenches, one or more metal gate material layers are formed in the gate trenches. In one embodiment, a metal layer **246** of a proper work function (referred to as a work function metal) and a conductive layer **248** are filled in the gate trenches. In one embodiment, the gate trenches in the device region **214** and overlay region **212** are deposited with a work function metal layer **246** and are then filled with the conductive material **248**, forming a gate electrode for a nFET. The work function metal **246** for the nFET is referred to as a n-metal. The n-metal includes a metal-based conductive material having a work function compatible to the nFET. For one example, the n-metal has a work function of about or less than about 4.2 eV. In one embodiment, the n-metal includes tantalum (Ta). In another embodiment, the n-metal includes titanium aluminum nitride (TiAlN). In other embodiments, the n-metal includes Ta, TiAl, TiAlN, or combinations thereof. The n-metal may include various metal-based film as a stack for optimized device performance and processing compatibility. The n-metal layer can be formed by a suitable process, such as PVD. The conductive material layer **248** may include aluminum, tungsten or other suitable metal. Then, a CMP process may be applied to remove the excessive work function metal and the conductive material. In one embodiment, the device region **214** includes both nFETs and pFETs. In this embodiment, the metal gates are formed for the nFETs and pFETs, respectively by a proper procedure. For example, after the removal of the silicon from the silicon gate stacks, the metal gates for the nFETs and the overlay mark are formed by a deposition for the n-metal layer, a deposition for the conductive layer, and a CMP process to remove the excessive n-metal layer and the conductive layer while the pFETS are protected by a patterned photoresist layer. Then the metal gates for pFETs are formed by a deposition for the p-metal layer, a deposition for the conductive layer, and a CMP process to remove the excessive p-metal layer and the conductive layer. Alternatively, a p-metal layer is deposited for the pFETs while the nFETs are protected by a patterned photoresist layer. A n-metal layer is deposited for the nFETs and the overlay mark while the pFETs are protected by a patterned photoresist layer. Then, a conductive layer is deposited to fill gate trenches for nFETs, pFETs and the overlay mark. A CMP process is applied to the substrate to remove the excessive portion of the n-metal layer, p-metal layer, and conductive layer, forming the metal gates for nFETs, pFETs and overlay mark.

The p-metal includes a metal-based conductive material having a work function compatible to the pFET. For one example, the p-metal has a work function of about 5.2 eV or greater. In one embodiment, the p-metal includes titanium nitride (TiN) or tantalum nitride (TaN). In other embodiments, the p-metal include TiN, tungsten nitride (WN), tantalum nitride (TaN), or combinations thereof. The p-metal may include various metal-based film as a stack for optimized device performance and processing compatibility. The p-metal layer can be formed by a suitable process, such as physical vapor deposition (PVD), CVD, ALD, PECVD, PEALD or spin-on metal. The conductive material thereafter substantially fills in the gate trench. The conductive material includes aluminum or tungsten according to various embodiments. The method to form the conductive material may include PVD, CVD, ALD, PECVD, PEALD or spin-on metal. Then, a CMP process may be applied to remove the excessive work function metal and the conductive material, forming the metal gate. Although the semiconductor structure **200** only illustrates one field-effect transistor in the device region **214**, a plurality of FETs and other devices can be formed in the device region. The present process to form the metal gates may have other alternative embodiment. For example, the metal gate for nFETs and pFETs may be formed by other sequence or other procedure.

In one embodiment, the metal gates may include a step to deposit a high-k dielectric material layer **244** the silicon oxide layer **220** in the gate trenches, and then a work function metal layer and a conductive layer are formed on the high-k dielectric material layer **244**. This process is referred to as a high-k last process. Alternatively, in the high-k last process, the silicon oxide layer **220** is first removed before forming the work function metal layer and conductive material layer. In this case, a new interfacial layer, such as silicon oxide, is first formed on the semiconductor substrate **210**, then the high-k dielectric material layer, work function metal layer and conductive material layer are formed in the corresponding gate trenches.

As described above, the overlay mark in the overlay region **212** including the gate stack **226** in the overlay region **212** is replaced by a metal gate. Particularly, the polysilicon in the gate stack **226** is replaced to form metal gates similar to the metal gate for the n-FETs in the device region **214** and formed in the same process to form the metal gates for the n-FETs. Therefore, the gate stacks in the overlay region **212** include the n-metal layer and the conductive material layer. In another example, the polysilicon gate stack **226** may be replaced to form metal gates similar to the metal gates for the p-FETs in the device region **214** and formed by the same process to form the metal gates for the p-FETs. In this case, the gate stacks in the overlay region **212** include the p-metal layer and the conductive material layer.

In another embodiment, the gate stacks for the overlay mark in the overlay region **212** remain as polysilicon gate stacks without replacement. In this case, the overlay region **212** is covered by a patterned mask layer such as a patterned photoresist layer or a patterned hard mask layer during the gate replacement to form metal gates for devices in the device region **214**, as illustrated in FIG. **9**.

Referring to FIG. **10**, the method **100** may proceed to step **114** by forming contact holes to electrical interconnection. In one embodiment, a contact etch stop layer (CESL) **250** is formed on the ILD layer **242** and another ILD layer **252** is formed on the CESL **250**. Then a photoresist layer (not shown) is coated on the semiconductor structure **200** in a lithography process, and a soft baking may be applied to the coated photoresist layer. Then a photomask (or mask) having a contact pattern is placed on the lithography exposure apparatus and the semiconductor structure **200** is secured on a wafer stage of the lithography exposure apparatus. Then the photomask is aligned to the semiconductor structure **200**